Planning with Large Language Models for Code Generation

Year: 2023 **|** *Citations: 211* **|** *Authors: Shun Zhang, Zhenfang Chen, Yikang Shen*

Abstract

Existing large language model-based code generation pipelines typically use beam search or sampling algorithms during the decoding process. Although the programs they generate achieve high token-matching-based scores, they often fail to compile or generate incorrect outputs. The main reason is that conventional Transformer decoding algorithms may not be the best choice for code generation. In this work, we propose a novel Transformer decoding algorithm, Planning-Guided Transformer Decoding (PG-TD), that uses a planning algorithm to do lookahead search and guide the Transformer to generate better programs. Specifically, instead of simply optimizing the likelihood of the generated sequences, the Transformer makes use of a planner to generate candidate programs and test them on public test cases. The Transformer can therefore make more informed decisions and generate tokens that will eventually lead to higher-quality programs. We also design a mechanism that shares information between the Transformer and the planner to make our algorithm computationally efficient. We empirically evaluate our framework with several large language models as backbones on public coding challenge benchmarks, showing that 1) it can generate programs that consistently achieve higher performance compared with competing baseline methods; 2) it enables controllable code generation, such as concise codes and highly-commented codes by optimizing modified objective.